August 19, 2011

# Choosing Assessments for Measuring Student Growth[1]

**Prepared for the Student Growth Measure Task Force[2]**
**Office of the State Superintendent of Education**
**Race to the Top**

Stanley Rabinowitz, Ph.D.
Edynn Sato, Ph.D.
Elizabeth Berkes, Ph.D.
WestEd

## Priority Grades and Subjects

| |
|---|
| **Grade 2, Reading** |
| **Grade 2, Mathematics** |
| **Grade 3, Reading (a fall to spring growth measure)** |
| **Grade 3, Mathematics (a fall to spring growth measure)** |
| **Grade 9, English/Language Arts** |
| **Algebra I** |
| **Geometry** |
| **Grades K-1, Reading** |
| **Grades K-1, Mathematics** |
| **Grades 6-8, Social Studies** |
| **Grades 6-8, Science** |
| **Kindergarten readiness (end of pre-K or beginning of K)** |

## Introduction

Reliable, valid, and bias-free measurement of student academic growth is important for all education stakeholders, particularly in the current high-stakes accountability environment. The guidelines presented in this document are intended to assist local educators in evaluating the technical adequacy of assessments used to measure student growth. Collectively, the guidelines present a systematic method for judging and documenting the quality of evidence that establishes the technical adequacy of assessments of student progress toward valued learning outcomes.

[1] Developed by WestEd with funding of and assistance by REL Mid-Atlantic

[2] This document has been created to assist public schools in the District of Columbia in their selection of assessment instruments to measure student academic growth. It may be applied equally to either the untested content areas and grades as part of educator effectiveness models or as supplemental measures in ELA, math and science, as already assessed and included in these models.

All assessments must meet sufficient standards of technical adequacy, regardless of the intended use. When measuring student growth is part of this intent, additional technical requirements must be met. High-quality evidence is required for making informed decisions about which assessment instruments fit the purposes, populations, and valued content, and have the technical adequacy needed to provide high-quality student achievement data.

No assessment instrument is perfect. In some cases good evidence may be available about one aspect of an instrument, but not available about another important aspect of that instrument. Or there may be cases in which an instrument is of high quality for one purpose, but the evidence shows that the instrument is weak for another. The evidentiary base may also vary between available instruments. In these cases it is important to have guidelines that allow decision-makers to weigh the sufficiency, quality, and adequacy of the evidence that is available for assessment instruments. What are the characteristics of appropriate assessments to measure student growth? What technical criteria should be applied? How do we know if the technical evidence about the assessment is sufficient, adequate, and of good quality to inform decisions? What indicators can be applied to help make a defensible selection decision?

In using these guidelines to determine the technical adequacy of a particular assessment instrument, decision-makers are likely to find that more evidence than is currently available is needed to make a final decision. DC schools are in a good position to collect and analyze this additional evidence during a pilot year. The technical guidelines presented here will be helpful in pinpointing the exact areas where more information is needed to make assessment selection decisions.

## Technical Criteria

Test consumers must distinguish between mere claims versus substantiated research findings regarding the technical adequacy of any assessment instrument. "Evidence" presented in technical documentation can range from assertions of findings (without support) to research summaries (without any evidence) to detailed descriptions of formal research-based technical evidence. It is important for consumers to consider the nature and source of the evidence when evaluating the technical adequacy of an assessment instrument under consideration.

A worksheet for documenting and evaluating the technical evidence associated with an instrument designed to assess student growth is available in Appendix A (page 7). Technical evidence can be found in such sources as technical and administration manuals, technical reports, Web postings, journal articles, and conference presentations. The criteria included on the worksheet are described below.

- Purpose: To ensure that the student performance data are appropriate for the specific required accountability intents, consumers of assessment data must be clear regarding the validated purpose(s) of the assessment instrument. This same principle applies for assessments of student growth. The validity of the student performance data may be compromised if the assessment is intended for one purpose and is used for another. The purpose of the assessment in terms of content and context must be documented, along with other general factors relevant to all assessments. If the assessment instrument is intended to be used to measure growth, it is important to know if the assessment was designed with that purpose in mind. If the instrument was not intended to measure growth, additional evidence must be collected to ensure that the instrument will be valid and reliable for that purpose.

- Population: To ensure that an assessment is valid and reliable for all populations of students, consumers need to verify that the assessment was field-tested with a population of students that reflects the same demographic background as the students with whom the assessment will be used. Factors to be taken into consideration include the full range of demographic variables such as ethnicity, cultural background, gender, socioeconomic background, education and language background, and particular disability characteristics. In addition to field testing, statistical and judgmental item reviews should be undertaken to ensure a lack of bias for all student populations.

- Content and Construct: Producers of assessments increase the validity and reliability of their instruments when they clearly articulate the range of skills and concepts to be assessed. Clearly determining the range of what students should know and be able to do in each domain of the assessment is particularly important when measuring growth. Proper alignment to the breadth and depth of this content should be demonstrated, particularly at all points of the intended score scale. Alignment studies are best performed by impartial third parties, though consumers can still judge the quality of in-house efforts. In addition, the validity and reliability of measurements can be affected by item type (e.g., multiple choice, constructed response). Consumers of assessment instruments that are intended to measure growth should be clear about which item types will provide the best student performance data for the target construct.

## Sufficiency, Quality, and Adequacy

Once the purpose, target population, and content of an assessment are determined to be an appropriate match for the consumer's needs, the next step is to evaluate the nature of the technical evidence provided—its sufficiency, quality, and adequacy. More specifically, consumers should consider the following:

- Information provided: Is the information an assertion, a summary, or a detailed description?

- Type of data: Are the data provided quantitative, qualitative, or both?

- Sufficiency: Is the information comprehensive (e.g., is quantitative information presented with supporting textual context and interpretation)?

- Quality: Does the method satisfy statistical assumptions? Is the method replicable? Is the outcome accurate (i.e., is there minimal or acceptable measurement error)? Is the outcome generalizable and/or broadly applicable?

- Adequacy: Is the information credible? Does the information directly support the evidence being evaluated?

## Demonstrating Technical Adequacy[3]

- Validity: Validity evidence establishes that the test measures what it purports to measure. An instrument must be validated for each intended use. While most assessment summaries describe the need for demonstrating validity, most provide specific evidence derived from a range of approaches, such as expert review of items against state standards and test specifications, alignment studies, reviews of p-values, and standard errors of measurement. Limitations of the field-test sample and their impact on the interpretation of the p-values may also be presented. Consumers may also look for cross-tabulation tables and Pearson correlation coefficients to show the relationship between student performance on the assessment and teacher ratings of student academic ability. In the case of assessments intended to measure growth, it is likely that strong evidence will not be available in all areas. Or the evidence may be limited to simple assertions that content experts have reviewed assessment items, or it may consist of quantitative results reported with little or no discussion of context or meaning. It is important for consumers to look for documented evidence in the following areas:

    - construct validity (the test is measuring the target skills and content)
    - criterion validity (the test predicts success as defined by the consumers and has the expected relationships with other measures of the same construct)
    - consequential validity (there is evidence that adverse consequences are minimal)
    - freedom from bias (there is evidence that the assessment is fair for all students)

- Reliability: All assessments must ensure consistency of measurement overall and at various points of the score scale. Typically both measures of internal consistency reliability (e.g., coefficient Alpha) and the standard error of measurement (SEM) are provided for each test form, along with discussion of the interpretation of these values with respect to the reliability of the assessment scores. Generally good

[3] For a definition of key assessment terms, see *A Glossary of Assessment Terms in Everyday Language*, http://www.ccsso.org/Documents/2006/Assessing_Students_with_Disabilities_Glossary_2006.pdf

reliability is indicated by a Cronbach's Alpha score that falls between $\alpha = .8$ and .9. It is important to look for evidence of the assessment's reliability in the following areas:

  - scale
  - internal consistency
  - split-half
  - scorer/hand-scoring
  - test-retest
  - alternate form
  - individual and group scores
  - classification consistency
  - generalizability

- Lack of bias: There should be evidence that the assessment under consideration is fair for all students. Ideally, both judgmental (e.g., bias review panels) and statistical (e.g., DIF analyses) approaches are used to determine whether items may be biased. Evidence in the following areas should be available about the instrument:

  - content
  - ethnicity
  - culture
  - linguistic
  - socioeconomic
  - geographic
  - students with disabilities
  - universal design

- Test administration procedures: In evaluating a potential assessment for use in measuring student growth, it is important to gather evidence about the availability of administrator training and supports for key administration procedures. Deviation from these procedures can have a serious effect on test validity and reliability. Information about the following procedures should be gathered:

  - assessment administration timing
  - scripts used to guide students
  - collection of non-cognitive information about students
  - distribution and collection of student assessment materials

- Scoring and reporting (interpretive guides): Scoring and reporting guides describe the types of scores generated and the structure of the score reports, and provide information on the meaning and the use of the data in the reports. Assessment consumers should look for evidence related to the following:

  - student level characteristics
  - NCLB subgroups
  - class
  - district
  - state
  - population
  - descriptions of standards setting

## Factors Specific to Assessments for Growth

In addition to ascertaining that an assessment is fair, reliable, and valid, further information is needed when selecting an assessment to measure student growth. Specifically, consumers need to determine whether the instrument is sensitive enough to measure true achievement gains versus measurement errors. This issue is especially relevant when dealing with high- or low-achieving students. Below we describe four attributes of an assessment that will increase the likelihood that it will be an appropriate tool to use to measure student growth. Many instruments will not fully meet these ideal conditions. Users must then decide if the evidence that is presented is sufficient, or if additional studies can be implemented in the pilot year to obtain more direct evidence of support.

- Multiple equated forms: Many aspects of assessments are “memorable,” particularly reading passages and mathematics word problems. Unless multiple forms are available to use in a pre-post design, some gains in student performance may be attributable to students’ memory of the past testing experience. The availability of alternate equated forms can address this concern, especially in the fall-to-spring model. This is often less of an issue in spring-to-spring approaches, since many assessments change forms from grade to grade. For spring- to-spring growth models, grade-specific forms must be placed on a comparable vertical scale, or studies need to be done on the vertical articulation of achievement standards.

- Recommended pre-post time frame: Time plays a key factor in student achievement. Unless sufficient instructional time elapses, we cannot reasonably attribute student gains in scores to instructional practices—the gains are more likely just errors of measurement. Test documentation should provide some guidance on what a reasonable instructional window is for measuring true improvement. (This issue is less likely a factor in fall-to-spring or spring-to-spring growth models; as the time frame decreases to quarterly or less, more problems exist.)

- Reliability at various points on the score scale: Testing time is not unlimited—test publishers must make decisions about how to select the numbers and types of items to fit a “reasonable” finite administration time. Often this requires selecting items near key decision points such as the proficiency cut score. This approach maximizes the reliability at that point. However, this means that other points of the score scale, particularly at the extremes, can have a significantly lower reliability. This concern is compounded when measuring growth; a measurement with a high degree of error at a single point cannot be expected to produce reliable growth estimates. Users should examine what are known as “conditional” standard errors to evaluate the reliability of score points across the scale, particularly at those points where large numbers of students are expected to fall. (This problem is ameliorated to a large extent if a Computer Adaptive Testing [CAT] model is used, since items are tailored to each student’s achievement level.)

- Evidence of gain score reliability: Related to the previous point, users should see if the publisher has evidence of the reliability of gains scores. Care should be taken to examine:

    - under what conditions the evidence was obtained;
    - what student populations were included in the studies; and
    - what time frame was used between the pre and post conditions.

**Pilot Studies**

As indicated earlier, no instrument will be perfect. In some cases, additional information may be required or desired before a final decision can be made as to whether the instrument has sufficient and adequate quality data to support its use. Fortunately there is a pilot year built into the schedule to allow for supplemental data collection, analysis, and review.

Technical studies require a certain degree of expertise to plan and implement. Should the school lack that expertise, the following strategies may be employed:

- Identify existing studies that have been performed in the areas where evidence may be lacking. Such studies may have been undertaken for the assessment under review with a different population group than exists at your school or may be on a test with similar characteristics as the one in question. Adapt the features of the existing study to fill the evidence gap identified by the review. For example, you may be able to determine such important study features such as:
    - defensible design features such as matched groups or treatment vs. control
    - sufficient sample size for reliable and valid results;
    - appropriate statistical analyses to determine effects.
    - user-friendly reporting formats for multiple audiences and constituencies.
- Request that the test publisher perform the necessary analyses; they may be willing as part of their services to the school or if they see the study having marketing value beyond the current situation.
- Team up with another school; you may be able not just to increase skill sets but sample size and other student demographic characteristics, allowing more sophisticated analyses.
- Partner with research organizations (universities, regional labs, comprehensive centers, etc.); these organizations typically have technical staff and a mission to support key educational reform initiatives.

With any of these approaches, careful attention to scope, plans, timelines, and budget will be essential for desired outcomes to be met in a timely fashion.

## **Appendix A:** Technical Evidence Worksheet

This worksheet guides assessment consumers through an evaluation of the context and content of an assessment instrument intended for measure student growth (Section I) and of the technical evidence presented to establish the assessment's validity, reliability, and freedom from bias (Sections II and III). Assessment producers can use this worksheet to verify that such information is presented in their assessment's documentation.

Test Name:

____________________________________________________________

Publisher:

____________________________________________________________

Year of Publication: _______________

## Section I

Instructions: First articulate your intended purposes, student population, and assessed content needs. Then, review the assessment's documentation (e.g., technical report, manuals) and determine whether this assessment is appropriate for your intents and needs.

## Section II

Instructions: Review the assessment's documentation and evaluate (i.e., 0-5, as defined in the column "Evaluation") the presence and presentation of the technical evidence listed below. To the degree possible, evaluate the assessment's "Specific Evidence."

| Criteria Cluster | Criterion | Specific Evidence[1] | Evaluation<br>0 = Unsure/unclear<br>1 = No information presented<br>2 = Evidence addressed in an assertion<br>3 = Evidence presented in a summary without data<br>4 = Evidence presented in a summary with data<br>5 = Evidence presented in a detailed description with data | Notes<br>(e.g., if Specific Evidence is not available, you may choose to note the level of detail at which evidence is available – Criteria Cluster, Criterion; questions or concerns; reference to other documentation that may further address a piece of technical evidence) |
|---|---|---|---|---|
| Validity | Field Testing | Field Test Sampling Design: Representativeness and Norming | | |
| | | Field Test Sampling Design: Currency (at least, dates documented) | | |
| | | Field Test Sampling Design: Randomization | | |
| | | Fidelity (link of test to stated purpose of the test) | | |
| | Design | Attrition of Persons (for Pre/Post Designs) | | |
| | | Test Blueprint | | |
| | | Scoring Rubric for OE Items: Construction and Validation | | |
| | | Accommodations | | |
| | Content | Content Alignment Studies | | |
| | | Expert judgments | | |
| | | p-values | | |
| | | Discrimination (Item-test Correlations) | | |
| | | Bias/DIF analysis | | |
| | | IRT/Item fit (ICC) | | |
| | | Distractor Analysis | | |
| | Construct | Factorial Validity (structural equation modeling) | | |
| | | Multi-Trait/Multi-Method | | |
| | | Equivalence/Comparability (construct the same regardless of examinee's ability) | | |
| | Criterion | Predictive validity - Validation to the Referent | | |
| | | Predictive validity - Individual and group scores | | |
| | | Concurrent validity - Validation to External Criteria | | |
| | | Concurrent validity - Validity of External Criteria | | |
| | | Concurrent validity - Individual and group scores | | |
| | Consequential | Evaluation of Testing Consequences | | |
| | | Individual and group scores | | |

[1] The specific evidence in this column is intended to represent an exhaustive list of technical evidence supporting sound tests and testing systems. Some of these elements may not be possible or appropriate for all types of tests.

| Criteria Cluster | Criterion | Specific Evidence[1] | Evaluation<br>0 = Unsure/unclear<br>1 = No information presented<br>2 = Evidence addressed in an assertion<br>3 = Evidence presented in a summary without data<br>4 = Evidence presented in a summary with data<br>5 = Evidence presented in a detailed description with data | Notes<br>(e.g., if Specific Evidence is not available, you may choose to note the level of detail at which evidence is available – Criteria Cluster, Criterion; questions or concerns; reference to other documentation that may further address a piece of technical evidence) |
|---|---|---|---|---|
| Validity, continued | Growth | Multiple equated forms | | |
| | | Recommended pre-post time frame | | |
| | | Reliability at various points of score scale | | |
| | | Gain score reliability | | |
| Reliability | Reliability: Single Administration | Scale | | |
| | | Internal Consistency | | |
| | | Split-half | | |
| | | Scorer/Hand-scoring | | |
| | Reliability: Multiple Administrations | Test-retest | | |
| | Reliability: Either Single or Multiple Administrations | Alternate form | | |
| | | Individual and group scores | | |
| | | Classification consistency | | |
| | | Generalizability | | |
| Freedom from Bias | Judgmental and Statistical (DIF) Reviews | Bias review panel | | |
| | | Content | | |
| | | Ethnicity | | |
| | | Cultural | | |
| | | Linguistic | | |
| | | Socio-economic | | |
| | | Geographic | | |
| | | Students with disabilities | | |
| | | Universal Design | | |

[1] The specific evidence in this column is intended to represent an exhaustive list of technical evidence supporting sound tests and testing systems. Some of these elements may not be possible or appropriate for all types of tests.

(*continued*)

| Criteria Cluster | Criterion | Specific Evidence[1] | Evaluation<br>0 = Unsure/unclear<br>1 = No information presented<br>2 = Evidence addressed in an assertion<br>3 = Evidence presented in a summary without data<br>4 = Evidence presented in a summary with data<br>5 = Evidence presented in a detailed description with data | Notes<br>(e.g., if Specific Evidence is not available, you may choose to note the level of detail at which evidence is available – Criteria Cluster, Criterion; questions or concerns; reference to other documentation that may further address a piece of technical evidence) |
|---|---|---|---|---|
| Testing System (Superordinate) Criteria | Form-Level Analyses | N | | |
| | | Central Tendency (Mean, Median, Mode) | | |
| | | Variation (Range, Variance, Standard Deviation) | | |
| | | Standard Error of Measurement | | |
| | | Bias | | |
| | | IRT fit (TCC) | | |
| | | Equating | | |
| | | Scaling | | |
| | Reporting | Student level | | |
| | | NCLB Subgroups | | |
| | | Class | | |
| | | District | | |
| | | State | | |
| | | Population | | |
| | | Description of Standards Setting: Methods, Participants, Group Size | | |
| | Report Format | Basic | | |
| | | Custom | | |

[1] The specific evidence in this column is intended to represent an exhaustive list of technical evidence supporting sound tests and testing systems. Some of these elements may not be possible or appropriate for all types of tests.

## Section III

Instructions: Once you've completed your evaluation of the assessment's technical evidence, consider the following:

1. The information related to the assessment's technical evidence consists mostly of:
   - □ Assertions
   - □ Summaries
   - □ Detailed descriptions

2. The evidence and related information provided in the assessment's documentation is (check all that apply):
   - □ Comprehensive (e.g., quantitative information is accompanied by supporting text and interpretations)
   - □ Accurate and directly supports the evidence being evaluated
   - □ Generalizable or broadly applicable
   - □ Credible

3. The data presented to support the technical evidence discussed in the assessment's documentation are mostly:
   - □ Quantitative
   - □ Qualitative
   - □ Both quantitative and qualitative
   - □ There are no data presented

4. There is enough evidence to start using the test.
   - □ Yes
   - □ Yes with reservations (reason): ________________________________________
   - □ No

5. If there is currently insufficient evidence, there is a plan to gather the evidence needed.
   - □ Yes (explanation): ____________________________________________
   - □ No